Bayesian Inference in Hidden Markov Random Fields for Binary Data Defined on Large Lattices

Friel, N.; Pettitt, A.N.; Reeves, R.; Wit, E.

Published in:
Journal of Computational and Graphical Statistics

DOI:
10.1198/jcgs.2009.06148

Supplemental materials for this article are available through the JCGS web page at http://www.amstat.org/publications.

Bayesian Inference in Hidden Markov Random Fields for Binary Data Defined on Large Lattices

N. FRIEL, A. N. PETTITT, R. REEVES, and E. WIT

Hidden Markov random fields represent a complex hierarchical model, where the hidden latent process is an undirected graphical structure. Performing inference for such models is difficult primarily because the likelihood of the hidden states is often unavailable. The main contribution of this article is to present approximate methods to calculate the likelihood for large lattices based on exact methods for smaller lattices. We introduce approximate likelihood methods by relaxing some of the dependencies in the latent model, and also by extending tractable approximations to the likelihood, the so-called pseudolikelihood approximations, for a large lattice partitioned into smaller sublattices. Results are presented based on simulated data as well as inference for the temporal-spatial structure of the interaction between up- and down-regulated states within the mitochondrial chromosome of the *Plasmodium falciparum* organism. Supplemental material for this article is available online.

Key Words: Autologistic model; Ising model; Latent variables; Markov chain Monte Carlo methods; Normalizing constant.

1. INTRODUCTION

This article is concerned with the problem of carrying out Bayesian inference for a hidden Markov random field model. This is an example of a general statistical problem of the following type: observed datum $\mathbf{y}$ masks or hides some unobserved latent or missing process $\mathbf{x}$. Denote all model parameters by θ. Interest may be in inference about parameters θ, or about the latent or missing datum $\mathbf{x}$. The marginal posterior distribution for θ, $p(\theta|\mathbf{y})$, is often intractable, but computation can often be simplified by including the hidden or missing datum $\mathbf{x}$ in the inference procedure. Widely studied examples of this setup

N. Friel is Associate Professor, School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Ireland (E-mail: *nial.friel@ucd.ie*). A. N. Pettitt is Professor, School of Mathematical Sciences, Queensland University of Technology, Australia. R. Reeves is Lecturer, School of Mathematical Sciences, Queensland University of Technology, Australia. E. Wit is Professor, Institute for Mathematics and Computer Science, University of Groningen, The Netherlands.

Journal of Computational and Graphical Statistics, Volume 18, Number 2, Pages 243–261
DOI: 10.1198/jcgs.2009.06148

include mixture models and hidden Markov models. In this article we consider the situation where the latent hidden variable $\mathbf{x}$ takes the form of a Markov random field (MRF). This problem is thus seen as one of performing inference for a hierarchical model, or more generally a directed graphical model, where a hyperprior is placed on the distribution of the hidden MRF. Besag, York, and Mollié (1991) presented an early analysis of this type of problem. In fact, the hugely influential work of Geman and Geman (1984) examined a similar problem in image analysis, but where the hidden Ising model had a known parameter value.

A major difficulty with this type of problem is that the likelihood $p(\mathbf{x}|\theta)$ of the hidden layer given the model parameters is often intractable, due to the difficulty of calculating its normalizing constant. Various approaches have been presented in the literature tackling this problem using Monte Carlo methods. Geyer and Thompson (1992) presented a Monte Carlo approach to estimating the normalizing constant $z(\theta)$. More recently, Gu and Zhu (2001) and Zhu, Gu, and Peterson (2007) have presented the stochastic approximation expectation algorithm to compute the maximum likelihood estimator for hidden Markov random field models. Briefly, their approach is to approximate first- and second-order partial derivatives of the log-likelihood $p(\mathbf{y}|\theta)$, using Monte Carlo averages. These are then employed in the context of a gradient-type optimization algorithm. Liang (2007) presented an alternative Monte Carlo-based approach. Here the normalizing constant $z(\theta)$ is viewed as a marginal distribution of the unnormalized distribution $q(\mathbf{x}, \theta)$. This approach illustrates how a kernel density estimate of $z(\theta)$ can be formulated based on Monte Carlo draws from $p(\mathbf{x}|\theta)$.

Nonsimulation-based methods have also been proposed in the literature to calculate the normalizing constant in an efficient manner (Pettitt, Friel, and Reeves 2003; Reeves and Pettitt 2004). The method presented in Pettitt, Friel, and Reeves (2003) involves calculating the normalizing constant for a lattice where each column in the lattice has two nearest column neighbors—the lattice can be viewed as being wrapped on a cylinder. Reeves and Pettitt (2004) presented an exact method, termed the *recursion method*, for calculating the normalizing constant for an unnormalized distribution expressible as a product of factors, of which the Ising model and related distributions are examples. This method is constrained, computationally, to relatively small lattices, where the smaller of the two dimensions of the lattice is no greater than 20 for a moderate number of the other dimension.

The main contribution of this article is to show how the recursion method can be extended in different ways to approximate normalizing constants for large lattices and allow likelihoods of the latent process to be approximated. The first approximation we propose, which we term the *reduced dependence approximation*, results from relaxing some of the dependencies in the latent model, so that $p(\mathbf{x}|\theta)$ is approximated as a product of factors each of which is defined on sublattices whose normalizing constant can be calculated via the recursion method. The method of pseudolikelihood introduced by Besag (1974), and partially ordered Markov models (Cressie and Davidson 1998), both approximate the likelihood as a product of tractable conditional distributions. The second approximation which we introduce results from extending pseudolikelihood to the case where the lattice is partitioned into blocks of sublattices, and where the likelihood of the overall lattice is approximated by a pseudolikelihood function defined on the blocks of sublattices. We term this

method *block pseudolikelihood* estimation. Performance of these new likelihood approximations is then illustrated in the context of inference for hidden Markov random fields for simulated and real data.

Other techniques exist to estimate the likelihood. Thermodynamic integration or path sampling (Gelman and Meng 1998) allows estimation of the log ratio of normalizing constants. In the context of hidden Markov fields this method has been employed by, for example, Green and Richardson (2002), Dryden, Scarr, and Taylor (2003), and Sebastiani and Sørbye (2002). The first two articles extend the problem to one of model selection where the number of hidden states is itself a parameter. All three articles, however, use an off-line approach to calculating the normalizing constant. In addition, a vast literature exists in the machine learning community presenting variational approximations to intractable MRF models. Recent work by, for example, Murray and Ghahramani (2004) outlined a variety of different approximation schemes in this context.

The article takes the following form. Section 2 introduces the main inference problem, illustrating the difficulties therein; Section 3 presents a review of different methods used to calculate the likelihood of $\mathbf{x}$ given β. Section 4 introduces the various large lattice likelihood approximations, which are then used in the inferential process. Section 5 presents a simulation study which is carried out to compare the different methods on simulated data. Section 6 describes an application of the methodology to real data involving gene expression levels from a time course microarray experiment for a genome in which the exact location of all the genes on the genome is known. Interest then concerns whether expression levels at a particular gene influence the expression level of neighboring genes. Finally, Section 7 presents a discussion of the various methods including some possible extensions of the methodology.

2. INFERENCE PROBLEM

2.1 HIDDEN MARKOV RANDOM FIELD MODELS

Suppose we are given an observed lattice of data values $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ where each value y_i is an observed value depending on some underlying discrete variable x_i from a lattice $\mathbf{x}$. We assume that conditional on $\mathbf{x}$ the y_i's are independent, so that

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p(y_i|x_i, \mu)$$

for some parameters $\mu(x_i)$. If the values of x_i in $\mathbf{x}$ are all uncorrelated, then $\mathbf{y}$ represents a sample from a mixture distribution. Here we are concerned with the situation where $\mathbf{x}$ is distributed as a Markov random field taking values $\{-1, 1\}$.

The autologistic model (Besag 1974) is an example of a first-order binary Markov random field, defined as follows:

$$p(\mathbf{x}|\beta) = \frac{q(\mathbf{x}|\beta)}{z(\beta)} = \frac{\exp(\beta_0 V_0(\mathbf{x}) + \beta_1 V_1(\mathbf{x}))}{z(\beta)}, \tag{2.1}$$

where

$$V_0(\mathbf{x}) = \sum_{i=1}^{n} x_i \qquad \text{and} \qquad V_1(\mathbf{x}) = \sum_{i \sim j} x_i x_j \tag{2.2}$$

and where we define $\beta = (\beta_0, \beta_1)$. The notation $i \sim j$ means that x_j is a neighbor of x_i, and each neighboring pair enters the summation only once. The statistic $V_1(\mathbf{x})$ is termed the energy function in statistical physics. Large positive values of β_1 lead to realizations of $\mathbf{x}$ having patches of -1's or $+1$'s. The parameter β_0 controls the relative abundance of -1's and $+1$'s. Positive or negative values of β_0 tend to encourage relatively more $+1$ or -1 states, respectively. When $\beta_0 = 0$, the autologistic model reduces to the Ising model. Finally, $z(\beta)$ corresponds to the normalizing constant, or in statistical physics terminology, the partition function:

$$z(\beta) = \int_{\mathbf{x}} \exp(\beta_0 V_1(\mathbf{x}) + \beta_1 V_1(\mathbf{x}))\nu(d\mathbf{x}), \tag{2.3}$$

where ν is a counting measure. Examples of studies where the autologistic model has been used include Augustin, Mugglestone, and Buckland (1996), Preisler (1993), and Wu and Huffer (1997). Throughout we assume that the distribution of states $\mathbf{x}$ is defined on a lattice of size $m \times m'$, where $n = mm'$, and where points are indexed from top to bottom in each column, and where columns are ordered from left to right. We assume in the following a first-order neighborhood model, with the neighbors of an interior point x_i denoted $\{x_{i-m}, x_{i-1}, x_{i+1}, x_{i+m}\}$. Along the edges of the lattice each point has either two or three neighbors. The full conditional distribution of x_i can then be written as

$$p(x_i|x_{\setminus i}, \beta) \propto \exp(\beta_0 x_i + \beta_1 x_i(x_{i-m} + x_{i-1} + x_{i+1} + x_{i+m})). \tag{2.4}$$

Here $x_{\setminus i}$ denotes the set $\mathbf{x}$ excluding the point x_i. Again this conditional distribution is modified along the edges of the lattice. Equivalence between the models formulated in (2.1) and (2.4) is given by the Hammersley–Clifford theorem (e.g., Besag 1974).

2.2 Inference for Hidden Markov Random Field Models

The primary problem of interest is to make inference about all unknown parameters conditional on the observed data $\mathbf{y}$; in other words, to evaluate the posterior distribution $p(\mathbf{x}, \mu, \beta|\mathbf{y})$. Assuming that β and μ are a priori independent allows the posterior to be formulated as

$$p(\mathbf{x}, \mu, \beta|\mathbf{y}) \propto \left\{ \prod_{i=1}^{n} p(y_i|x_i, \mu) \right\} p(\mathbf{x}|\beta)\pi_\beta(\beta)\pi_\mu(\mu) \tag{2.5}$$

where $\pi_\beta(\cdot)$ and $\pi_\mu(\cdot)$ are prior distributions for β and μ, respectively.

To generate samples from the posterior we proceed in standard fashion by running an MCMC (Markov chain Monte Carlo) sampler drawing parameter values from their full conditional distribution. We describe the algorithm below:

Step 1: Update each x_i in turn by Gibbs sampling from

$$p(x_i|\mathbf{x}_{\setminus i}, \mathbf{y}, \beta, \mu) \propto p(y_i|x_i, \mu)\, p\big(x_i|x_{N(i)}, \beta\big). \tag{2.6}$$

Step 2: Update μ: Carry out a Metropolis–Hastings update of μ from the full conditional distribution

$$p(\mu|\mathbf{x}, \mathbf{y}, \beta) \propto \left\{\prod_{i=1}^{n} p(y_i|x_i, \mu)\right\}\pi_\mu(\mu).$$

Step 3: Update β: Carry out a Metropolis–Hasting update of β from the full conditional distribution

$$p(\beta|\mathbf{x}, \mu, \mathbf{y}) \propto p(\mathbf{x}|\beta)\pi_\beta(\beta).$$

Each of Steps 1 and 2 poses no major problems—it is straightforward to design samplers to sample from the respective full conditionals. However, Step 3 is problematic. Here the probability of $\mathbf{x}$ given β involves knowledge of the normalizing constant of the MRF. Examining (2.3), it is clear that this involves a sum of $2^{mm'}$ terms, which is infeasible even for very small lattice sizes. We focus on this issue in the next section.

3. REVIEW OF LIKELIHOOD TECHNIQUES FOR MRFS

This section reviews different methods to compute or estimate $p(\mathbf{x}|\beta)$.

3.1 PSEUDOLIKELIHOOD ESTIMATION

The likelihood of $\mathbf{x}$ given β, namely $p(\mathbf{x}|\beta)$, carries a severe computational load, because it requires the calculation of a normalizing constant $z(\beta)$. The most common approach to overcome this considerable computational problem is to approximate the likelihood using pseudolikelihood, first presented by Besag (1975). Here $p(\mathbf{x}|\beta)$ is approximated by a product of the full conditional probabilities for each lattice point:

$$p(\mathbf{x}|\beta) \approx \prod_{i=1}^{n} p(x_i|x_{\setminus i}, \beta). \tag{3.1}$$

Now by the property of Markov random fields, each term in the product only involves nearest-neighbor adjacencies, and so the normalizing constant of each full conditional is trivial to compute. We note that the right side of (3.1) is not generally normalized correctly with respect to $\mathbf{x}$. This method has been employed in a wide variety of settings. In particular, it has been used in the current context of hidden Markov random fields by, for example, Besag, York, and Mollié (1991), Rydén and Titterington (1998), and Heikkinen and Högmander (1994).

3.2 PARTIALLY ORDERED MARKOV MODELS

Partially ordered Markov models (POMMs) (Cressie and Davidson 1998) are a generalization of the Markov chain to a directed acyclic graph (DAG), and generalize Markov mesh models (MMMs) (Abend, Harley, and Kanal 1965). They have the advantage that the likelihood is directly available as a product of conditional probabilities, without the

need for computing a normalizing constant. Whereas there is an equivalent Markov random field for any specific POMM, only a subset of Markov random fields are expressible as POMMs. For other Markov random fields, it may be possible to find an approximating POMM that gives approximately the same probability for any particular lattice. Goutsias (1991) presented an approach for finding a Markov mesh model, which he termed a mutually compatible Gibbs random field, to approximate a general Markov random field, or as he termed it, a general Gibbs random field.

In general this is an optimization problem to maximize the similarity between two distributions, in which the free parameters of the POMM approximation are the association parameter and the parentage structure. However, this does not necessarily preserve the interpretation of the association parameter, which is desirable in our application.

Requiring the parameter to be equal in each model necessitates finding an approximating POMM of the form

$$p(\mathbf{x}|\beta) \approx \prod_{i=1}^{n} p(x_i|\,\mathrm{pa}(x_i), \beta), \tag{3.2}$$

where $\mathrm{pa}(x_i) = \{x_{i+1}, x_{m+i}\}$ denotes the parents of the point x_i. Along the right column of the lattice, $\mathrm{pa}(x_i) = x_{i+1}$ and along the bottom row of the lattice, $\mathrm{pa}(x_i) = x_{m+i}$, except for point x_n which has no parents.

3.3 Generalized Recursions

Generalized recursions for computing the normalizing constant of general factorizable models such as autologistic and Potts models have been proposed by Reeves and Pettitt (2004), generalizing a result known for hidden Markov models (e.g., Zucchini and Guttorp 1991; Scott 2002). This method applies to autologistic lattices with a small number of rows, up to about 20, and is based on an algebraic simplification due to the reduction in dependence arising from the Markov property. It applies to unnormalized likelihoods that can be expressed as a product of factors, each of which is dependent on only a subset of the lattice sites. We can write $q(\mathbf{x}|\beta)$ in factorizable form as

$$q(\mathbf{x}|\beta) = \prod_{i=1}^{n} q_i(\mathbf{x}_i|\beta),$$

where each factor q_i depends on a subset $\mathbf{x}_i$ of $\mathbf{x}$ comprising the points $x_i, x_{i+1}, \ldots, x_{i+m}$, where m is defined to be the *lag* of the model. We may define each factor as

$$q_i(\mathbf{x}_i, \beta) = \exp\{\beta_0 x_i + \beta_1 x_i (x_{i+1} + x_{i+m})\} \tag{3.3}$$

for all i, except when i corresponds to a lattice point on the last row or last column, in which case x_{i+1} or x_{i+m}, respectively, drops out of (3.3).

As a result of this factorization, the summation for the normalizing constant

$$z(\beta) = \sum_{\mathbf{x}} \prod_{i=1}^{n} q_i(\mathbf{x}_i|\beta)$$

can be represented as

$$z(\beta) = \sum_{x_n} q_n(\mathbf{x}_n|\beta) \sum_{x_{n-1}} q_{n-1}(\mathbf{x}_{n-1}|\beta) \cdots \sum_{x_1} q_1(\mathbf{x}_1|\beta)$$

which can be computed much more efficiently than the straightforward summation over the 2^n possible lattice realizations. Full details of a recursive algorithm to compute the above can be found in Reeves and Pettitt (2004). The reader is also referred to Jordan (2004) where the same problem is addressed but from a more graph-theoretic perspective.

The minimum lag representation for an autologistic lattice with a first-order neighborhood occurs for r given by the smaller of the number of rows or columns in the lattice. Identifying the number of rows with the smaller dimension of the lattice, the computation time increases by a factor of 2 for each additional row, but linearly for additional columns.

4. APPROXIMATING LIKELIHOODS FOR LARGER LATTICES

In this section we introduce new methods to approximate the likelihood $p(\mathbf{x}|\beta)$ by showing how the recursion method can be exploited in very natural ways to estimate $z(\beta)$ for larger lattice sizes.

4.1 REDUCED DEPENDENCE APPROXIMATION

Define the vector of states in row i as $\mathbf{r}_i$. Writing the distribution of the lattice in terms of the $\mathbf{r}_i$'s and using the Markov property gives

$$p(\mathbf{x}|\beta) = p(\mathbf{r}_{m-m_1+1}, \ldots, \mathbf{r}_m|\beta) \prod_{i=1}^{m-m_1} p(\mathbf{r}_i|\mathbf{r}_{i+1}, \beta) \tag{4.1}$$

for some number $m_1 < m$. It holds that

$$p(\mathbf{r}_i|\mathbf{r}_{i+1}, \beta) = \frac{p(\mathbf{r}_1, \ldots, \mathbf{r}_{i-1}, \mathbf{r}_i|\mathbf{r}_{i+1}, \beta)}{p(\mathbf{r}_1, \ldots, \mathbf{r}_{i-1}|\mathbf{r}_i, \beta)} \tag{4.2}$$

for any realized values of the sublattice comprising the first $i-1$ rows, $(\mathbf{r}_1, \ldots, \mathbf{r}_{i-1})$. We can rewrite (4.2), again using the Markov property, as

$$p(\mathbf{r}_i|\mathbf{r}_{i+1}, \beta) = \frac{p(\mathbf{r}_1, \ldots, \mathbf{r}_{i-m_1-1}|r_{i+1}, \beta)}{p(\mathbf{r}_1, \ldots, \mathbf{r}_{i-m_1-1}|r_i, \beta)} \frac{p(\mathbf{r}_{i-m_1}, \ldots, \mathbf{r}_i|\mathbf{r}_{i+1}, \mathbf{r}_{i-m_1-1}, \beta)}{p(\mathbf{r}_{i-m_1}, \ldots, \mathbf{r}_{i-1}|\mathbf{r}_i, \mathbf{r}_{i-m_1-1}, \beta)}. \tag{4.3}$$

We introduce an approximation to (4.3) by assuming, in the numerator and denominator, that the sublattices $(\mathbf{r}_{i-m_1}, \ldots, \mathbf{r}_i)$ and $(\mathbf{r}_{i-m_1}, \ldots, \mathbf{r}_{i-1})$, respectively, are independent of their neighboring rows and that the sublattice $(\mathbf{r}_1, \ldots, \mathbf{r}_{i-m_1-1})$ is independent of r_i and r_{i+1}, leaving

$$p(\mathbf{r}_i|\mathbf{r}_{i+1}, \beta) \approx \frac{p(\mathbf{r}_{i-m_1}, \ldots, \mathbf{r}_i|\beta)}{p(\mathbf{r}_{i-m_1}, \ldots, \mathbf{r}_{i-1}|\beta)}. \tag{4.4}$$

Notice that each probability in the numerator and denominator on the right side of (4.4) is defined on a sublattice of dimension $(m_1+1) \times n$ and $m_1 \times n$, respectively. Therefore the probabilities on the right side of (4.4) can be computed exactly by the generalized recursion method provided $m_1 \leq 20$. Substituting (4.4) for each factor in (4.1) leads to

$$p(\mathbf{x}|\beta) \approx \frac{\exp(\beta_0 V_0(\mathbf{x}) + \beta_1 V_1(\mathbf{x}))(z_{m_1 \times n}(\beta))^{m-m_1-1}}{(z_{(m_1+1)\times n}(\beta))^{m-m_1}}, \tag{4.5}$$

where we now adopt the notation $z_{(m_1+1)\times n}(\beta)$ for the normalizing constant of an $(m_1+1) \times n$ sublattice. We expect that as β_1 increases, so too would the number m_1 of rows

needed for a good approximation, due to increasing correlation between lattice points a distance m_1 apart. From a computational viewpoint (4.5) shows that each time β is updated, it suffices to calculate two tractable low-dimension normalizing constants, one for an $m_1 \times n$ lattice and another for an $(m_1 + 1) \times n$ lattice, instead of one intractable high-dimension normalizing constant. We term this approximation the *reduced dependence approximation* (RDA).

Note that a similar idea has been proposed by Stein, Chi, and Welty (2004) to approximate Gaussian likelihoods for large spatial datasets. The idea above is similar in spirit to the pseudolikelihood estimator or more generally to composite likelihoods (Heagerty and Lele 1998; Cox and Reid 2004), because here an intractable likelihood is approximated as a product of smaller tractable factors.

In effect the reduced dependence approximation to the true likelihood $p(\mathbf{x}|\beta)$ arises by estimating the true normalizing constant $z(\beta)$ as

$$\bar{z}_{m_1}(\beta) = \frac{(z_{m_1 \times n}(\beta))^{m-m_1+1}}{(z_{m_1-1 \times n}(\beta))^{m-m_1}}.$$

It is possible to calculate $z(\beta)$ exactly for a 20×20 lattice. In Figure 1 we compare the true normalizing constant, $z(\beta)$, to the approximation, $\bar{z}_{m_1}(\beta)$, where $\beta_0 = 0$ and $\beta_1 = 0.2, 0.3, 0.35, 0.4$, for $m_1 = 3, 4, \ldots, 20$. This plot shows that as β_1 increases, the value of m_1 needed for $\bar{z}_{m_1}(\beta)$ to approximate $z(\beta)$ also increases. Focusing on the choice $\beta_1 = 0.4$, notice that for $m_1 \leq 6$, the ratio of true to approximation drops below 0.95 whereas, for example, when $m_1 = 10$ the ratio equals 0.998. In Figure 2 we display values of $\bar{z}_{m_1}(\beta)$ for a lattice of dimension 50×50 for parameter values $\beta = (0, 0.4)$, for $m_1 = 3, 4, \ldots, 16$.

It is known in general that the log of the normalizing constant is a convex function in the parameters β (Jordan 2004). Notice here that the approximation to the normalizing constant appears convex as a function of m_1. This fact would be useful to help to correct for the approximation. Furthermore notice that the approximation appears to be an underestimate.

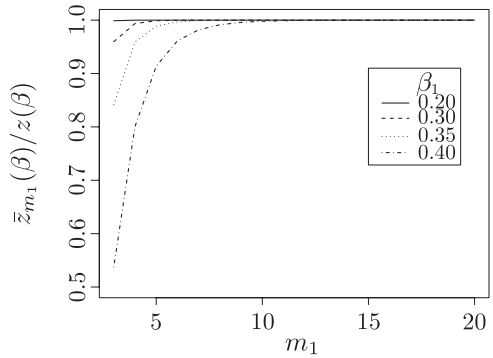

Figure 1. Ratio of the approximation to the normalizing constant, $\bar{z}_{m_1}(\beta)$, to the true normalizing constant, $z(\beta)$, for a lattice of dimension 20×20 for different β parameter values for $m_1 = 3, 4, \ldots, 20$.

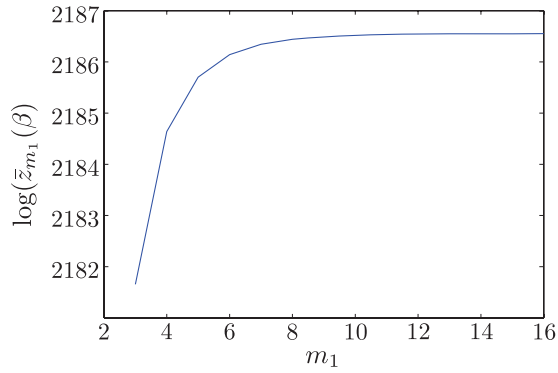

Figure 2. Approximation of log normalizing constant, $\bar{z}_{m_1}(\beta)$ for a lattice of dimension 50×50 for parameter value $\beta = [0, 0.4]$ for $m_1 = 3, 4, \ldots, 16$.

4.2 BLOCK PSEUDOLIKELIHOOD ESTIMATION

Recall from Section 3.1 that the pseudolikelihood estimate of $p(\mathbf{x}|\beta)$ is obtained as a product of the full conditional distribution of each lattice point. This approach can be extended by considering the full conditional distributions of blocks of lattice locations, conditional upon the rest of the lattice,

$$p(\mathbf{x}|\beta) \approx \prod_{l=1}^{L} p(\mathbf{x}_l|\mathbf{x}_{\setminus l}, \beta), \tag{4.6}$$

where the lattice is divided into L sublattices, and the notation $\mathbf{x}_{\setminus l}$ refers to those lattice locations outside sublattice $\mathbf{x}_l$. This differs slightly from the generalized pseudolikelihoods of Huang and Ogata (2002), which are based on overlapping blocks for each lattice location.

In the case of the autologistic model, for example, the full conditional for the sublattice is found by picking out all the terms of (2.1) which involve the sublattice $\mathbf{x}_l$,

$$p(\mathbf{x}_l|\beta, \mathbf{x}_{\setminus l}) = \frac{1}{z(\beta, \mathbf{x}_{\setminus l})} \exp(\beta_0 V_0(\mathbf{x}_l) + \beta_1 V_1(\mathbf{x}_l) + \beta_1 V_{neighs}(\mathbf{x}_l, \mathbf{x}_{\setminus l})), \tag{4.7}$$

where $V_{neighs}(\mathbf{x}_l, \mathbf{x}_{\setminus l})$ includes all the interaction terms between lattice locations in $\mathbf{x}_l$ and its boundary neighbors. This is an autologistic model on the sublattice, conditioned on the boundary values of all the neighboring sublattices. Because of the conditioning, the normalizing constant is dependent on the boundary values of the neighboring sublattices. However, the normalizing constants can be readily computed by the recursion method for reasonably sized sublattices.

Note that in a similar fashion we can extend the POMM approximation in Section 3.2 to define a block POMM approximation. Here the block POMM approximation differs from the block pseudolikelihood approximation in (4.7) by including only interaction terms between the sublattice $\mathbf{x}_l$ and two of its boundaries, to the bottom and left. We have explored using this approximation in an earlier draft of this article. In general it did not perform as well as the block pseudolikelihood approximation.

4.3 CLASSES OF APPROXIMATIONS

We have considered two broad classes of approximations to the autologistic distribution, namely, (i) a deterministic approximation to the normalizing constant and (ii) an approximation to the likelihood involving a product of normalized distributions of moderate to low dimension relative to the overall joint distribution.

In the first case, with the deterministic approximation $\bar{z}_{m_1}(\beta)$, the effect is equivalent to replacing the prior $\pi_\beta(\beta)$ in (2.5) by $\pi_\beta(\beta)\bar{z}_{m_1}(\beta)/z(\beta)$. The exploratory numerical illustration for the RDA (see Figure 1) suggests that the approximation provides values of $\bar{z}(\beta)/z(\beta)$ very close to 1.

In the second case, the pseudolikelihood or block pseudolikelihood approximation replaces $p(\mathbf{x}|\beta)$ in Step 3 of the algorithm. In the case of the block pseudolikelihood approximation, the use of the approximation is equivalent to the prior being changed to

$$\pi_\beta(\beta)\frac{\prod_{l=1}^{L} p(\mathbf{x}_l|\mathbf{x}_{\setminus l},\beta)}{p(\mathbf{x}|\beta)} \tag{4.8}$$

which possibly changes at each sweep or generation of the lattice values. The pseudolikelihood approximation results when the sublattice $\mathbf{x}_l$ reduces to a lattice point. How close $\prod_{l=1}^{L} p(\mathbf{x}_l|\mathbf{x}_{\setminus l},\beta)/p(\mathbf{x}|\beta)$ is to 1 depends on how large the sublattice $\mathbf{x}_l$ is. For the case of the pseudolikelihood estimator, this fraction often differs greatly from 1. Further, there is no guarantee that the resulting MCMC algorithm is stationary for the distribution defined in (2.5), or that the chain is positive recurrent. A positive recurrent chain would result if the autologistic distribution were replaced in Steps 1 and 3 of the algorithm by the pseudolikelihood or block pseudolikelihood model, and not just in Step 3, which is used to update β.

In all of these approximate methods for approximating the likelihood of a Markov random field, we are interested in their performance embedded within Markov chain Monte Carlo methods for inference for hierarchical models involving such models. In general we are interested in the computational efficiency of the resulting Markov chains, and the trade-off between computational and statistical efficiency in estimates derived from posteriors. We are also interested in whether these approximations introduce any discernible bias or additional variability into posterior distributions.

5. SIMULATION STUDY

Data were generated by gathering samples of realizations from autologistic models with distribution $p(\mathbf{x}|\beta)$. For each hidden data point x_{ij}, Gaussian noise with mean μ_{-1} or μ_1 and common variance of 1 was added conditional on $x_{ij} = -1$ or 1, respectively, resulting in an observed data point y_{ij}. We assume that the variance of the noise is known. Data of size 50×50 were generated for various combinations of β and $\mu = (\mu_{-1}, \mu_1)$. In total, 50 independent datasets were generated for each of 12 combinations of parameters. Each of the three estimation methods was applied to each of these 600 datasets, as part of an MCMC method to recover posterior distributions for each of the unknown autologistic and normal parameters, resulting in 1800 MCMC runs. Prior values for μ were distributed

uniformly from the set $\{(\mu_{-1}, \mu_1) | -5 \leq \mu_{-1} \leq 5, \mu_{-1} \leq \mu_1 \leq 5\}$. The prior for β was a flat Gaussian zero-mean prior.

The inference procedure was iterated as follows:

1. β was updated using a Metropolis–Hastings update from its full conditional distribution, using either (4.5), (4.6), or (3.1) to approximate $p(\mathbf{x}|\beta)$.
2. μ parameters were updated using a Metropolis–Hastings algorithm from their full conditional distributions.
3. Each point x_{ij}, in turn, was sampled from its Gibbs distribution; see (2.6).

When estimating the likelihood of $p(\mathbf{x}|\beta)$ for the RDA method we used the value $m_1 = 10$. The block pseudolikelihood method estimated the likelihood $p(\mathbf{x}|\beta)$ by splitting the lattice into 10×50 blocks.

All three approximate likelihood methods differ in terms of computation speed. The results were generated using a C program running under Linux on a Pentium IV 2.8 GHz processor with 512 Mb of RAM. The RDA method took 0.2 sec per full sweep of all parameters. Pseudolikelihood, by comparison, took 0.02 sec. The block pseudolikelihood method was slower overall, taking 0.6 sec per full sweep. Each method was run for the same computational time, resulting in a chain of length 30,000 iterations for RDA, 10,000 iterations for block pseudolikelihood, and 300,000 for pseudolikelihood.

In Tables 1 and 2 we present the average of the absolute bias of the posterior mean and the standard error of the posterior mean for each parameter combination. In Table 1, where $\mu = (-0.5, 0.5)$, the absolute bias of the posterior means is smaller for RDA and block pseudolikelihood compared to pseudolikelihood in all but one of 24 situtations, namely, for parameter μ_{-1}, when $\beta = (0.05, 0.3)$. In every other situation, both RDA and block pseudolikelihood lead to a smaller average absolute bias of the posterior mean. This is particularly true for cases when $\beta_1 = 0.4$ and $\beta_0 = 0, 0.05$, or 0.1.

The situation in Table 2, where $\mu = (-0.3, 0.3)$, represents a considerably more challenging scenario. The distance between the means of the noise distributions is 0.6 and is quite small relative to their common unit variance. The average absolute biases of the posterior means of the RDA and block pseudolikelihood methods are comparable and much less than the average absolute bias of the pseudolikelihood method for this scenario. For every parameter setting, RDA and block pseudolikelihood resulted in considerably smaller absolute biases. In general, there were not big differences between estimates from RDA and block pseudolikelihood methods. However, RDA is perhaps to be recommended, because it is computationally faster.

These results give very strong evidence that using RDA and block pseudolikelihood can lead to considerable improvement in parameter estimation compared to estimation via the pseudolikelihood method. These results are consistent with other studies which have outlined the shortcomings of the pseudolikelihood method, for example, Dormann (2007) and Sherman, Apanasovich, and Carroll (2006). Note also that the RDA method has been used in a similar context of hidden Markov random fields, but where inference is carried out using variational methods (McGrory et al. 2008). This article also concludes that RDA gives improved performance with respect to the pseudolikelihood method.

Table 1. Average absolute bias of the posterior mean (and standard error of the posterior mean) for 50 samples of 50×50 lattices, where the underlying hidden MRF is an autologistic model with given parameter specifications, where $\mu = (-0.5, 0.5)$.

	β_0	β_1	μ_{-1}	μ_1	True values
RDA	0.026 (0.006)	0.030 (0.006)	0.080 (0.015)	0.076 (0.014)	$\beta = (0, 0.3)$
Block pseudo	0.025 (0.006)	0.028 (0.004)	0.066 (0.014)	0.084 (0.015)	
Pseudo	0.032 (0.007)	0.033 (0.005)	0.078 (0.016)	0.091 (0.016)	
RDA	0.029 (0.006)	0.034 (0.007)	0.097 (0.017)	0.054 (0.010)	$\beta = (0.05, 0.3)$
Block pseudo	0.035 (0.006)	0.033 (0.006)	0.119 (0.019)	0.068 (0.011)	
Pseudo	0.036 (0.011)	0.047 (0.019)	0.091 (0.025)	0.112 (0.027)	
RDA	0.081 (0.015)	0.062 (0.011)	0.126 (0.024)	0.065 (0.012)	$\beta = (0.1, 0.3)$
Block pseudo	0.087 (0.019)	0.052 (0.009)	0.115 (0.023)	0.055 (0.013)	
Pseudo	0.104 (0.024)	0.052 (0.010)	0.150 (0.027)	0.147 (0.026)	
RDA	0.006 (0.001)	0.018 (0.003)	0.045 (0.008)	0.046 (0.008)	$\beta = (0, 0.4)$
Block pseudo	0.005 (0.001)	0.019 (0.003)	0.047 (0.008)	0.049 (0.009)	
Pseudo	0.011 (0.003)	0.057 (0.006)	0.062 (0.011)	0.107 (0.012)	
RDA	0.047 (0.014)	0.053 (0.009)	0.163 (0.027)	0.040 (0.007)	$\beta = (0.05, 0.4)$
Block pseudo	0.055 (0.010)	0.054 (0.010)	0.175 (0.032)	0.056 (0.010)	
Pseudo	0.270 (0.031)	0.210 (0.026)	0.334 (0.054)	0.607 (0.061)	
RDA	0.150 (0.038)	0.126 (0.027)	0.321 (0.058)	0.079 (0.018)	$\beta = (0.1, 0.4)$
Block pseudo	0.136 (0.028)	0.129 (0.026)	0.340 (0.058)	0.075 (0.017)	
Pseudo	0.545 (0.088)	0.359 (0.067)	0.508 (0.066)	0.729 (0.085)	

Table 2. Average absolute bias of the posterior mean (and standard error of the posterior mean) for 50 samples of 50×50 lattices, where the underlying hidden MRF is an autologistic model with given parameter specifications, where $\mu = (-0.3, 0.3)$.

	β_0	β_1	μ_{-1}	μ_1	True values
RDA	0.148 (0.032)	0.137 (0.023)	0.220 (0.037)	0.213 (0.041)	$\beta = (0, 0.3)$
Block pseudo	0.230 (0.040)	0.106 (0.015)	0.186 (0.035)	0.226 (0.039)	
Pseudo	0.380 (0.045)	0.419 (0.036)	0.632 (0.080)	0.460 (0.074)	
RDA	0.130 (0.028)	0.128 (0.026)	0.228 (0.039)	0.115 (0.023)	$\beta = (0.05, 0.3)$
Block pseudo	0.164 (0.044)	0.096 (0.018)	0.230 (0.043)	0.163 (0.028)	
Pseudo	0.504 (0.063)	0.445 (0.034)	0.487 (0.090)	0.895 (0.085)	
RDA	0.201 (0.039)	0.170 (0.028)	0.259 (0.044)	0.099 (0.020)	$\beta = (0.1, 0.3)$
Block pseudo	0.262 (0.005)	0.146 (0.001)	0.300 (0.086)	0.166 (0.062)	
Pseudo	0.577 (0.098)	0.496 (0.084)	0.564 (0.099)	0.889 (0.120)	
RDA	0.034 (0.011)	0.048 (0.009)	0.091 (0.018)	0.101 (0.019)	$\beta = (0, 0.4)$
Block pseudo	0.044 (0.011)	0.059 (0.013)	0.090 (0.017)	0.122 (0.023)	
Pseudo	0.387 (0.056)	0.266 (0.038)	0.516 (0.072)	0.578 (0.109)	
RDA	0.156 (0.031)	0.190 (0.034)	0.239 (0.044)	0.081 (0.019)	$\beta = (0.05, 0.4)$
Block pseudo	0.272 (0.042)	0.219 (0.025)	0.294 (0.055)	0.151 (0.024)	
Pseudo	0.582 (0.076)	0.633 (0.067)	0.745 (0.101)	0.754 (0.096)	
RDA	0.225 (0.038)	0.176 (0.029)	0.346 (0.053)	0.130 (0.022)	$\beta = (0.1, 0.4)$
Block pseudo	0.331 (0.053)	0.229 (0.035)	0.401 (0.054)	0.186 (0.028)	
Pseudo	0.720 (0.118)	0.567 (0.103)	0.700 (0.120)	0.881 (0.123)	

5.1 EXACT METHODS

Several recent approaches have appeared in the literature which allow inference for hidden Markov random fields by circumventing the problem of calculating intractable normalizing constants: an auxiliary variable method (Møller et al. 2006), and the exchange algorithm and developments (Murray 2007). We give a short review of these ideas.

5.1.1 Auxiliary Variable Method

Define an auxiliary variable $\mathbf{z}$ defined on the same state space as $\mathbf{x}$ with conditional density $f(\mathbf{z}|\beta, \mathbf{x})$. Define an augmented posterior distribution

$$p(\mathbf{x}, \mu, \beta, \mathbf{z}|\mathbf{y}) \propto \left\{ \prod_{i=1}^{n} p(y_i|x_i, \mu) \right\} f(\mathbf{z}|\beta, \mathbf{x})\, p(\mathbf{x}|\beta)\pi_\beta(\beta)\pi_\mu(\mu).$$

The actual distribution of interest is found by marginalizing over the auxiliary variable $\mathbf{x}$. In terms of an MCMC implementation, following Section 2.2, the only change to the algorithm is how β and $\mathbf{z}$ are updated. The key innovation in Møller et al. (2006) is to update these variables in a single step as follows. Suppose that the Markov chain is currently visiting β and $\mathbf{z}$. First β^* is proposed from a density $p(\beta^*|\beta)$. Then auxiliary variable $\mathbf{z}^*$ is proposed from the same likelihood model as $\mathbf{x}$, but depending on the proposed value β^*,

$$p(\mathbf{z}^*|\beta^*, \beta, \mathbf{z}) = \frac{q(\mathbf{z}^*|\beta^*)}{z(\beta^*)}. \tag{5.1}$$

The Metropolis–Hastings ratio for this joint update appears as

$$\frac{f(\mathbf{z}^*|\beta^*, \mathbf{x})q(\mathbf{x}|\beta^*)\pi_\beta(\beta^*)\, p(\beta|\beta^*)q(\mathbf{z}|\beta)}{f(\mathbf{z}|\beta, \mathbf{x})q(\mathbf{x}|\beta)\pi_\beta(\beta)\, p(\beta^*|\beta)q(\mathbf{z}^*|\beta^*)}.$$

But crucially the normalizing constants for each of $q(\mathbf{x}|\beta^*)$, $q(\mathbf{z}|\beta)$, $q(\mathbf{x}|\beta)$, and $q(\mathbf{z}^*|\beta^*)$ which appear in the numerator and denominator cancel above and below. A key step in the above scheme is the ability to sample $\mathbf{z}$ from the autologistic model (5.1). In this case perfect sampling is possible (Propp and Wilson 1996) and here we employ an algorithm based on a partial ordering of the states of the lattice.

5.1.2 Exchange Algorithm

The exchange algorithm also makes use of an auxiliary variable on the support of the data, as above, which however depends only on an additional auxiliary variable β' on the support of the parameter, which in turn depends on the parameter β. The augmented posterior distribution is then given by

$$p(\mathbf{x}, \mu, \beta, \beta', \mathbf{z}|\mathbf{y}) \propto \left\{ \prod_{i=1}^{n} p(y_i|x_i, \mu) \right\} p(\mathbf{z}|\beta')\, p(\beta'|\beta)\, p(\mathbf{x}|\beta)\pi_\beta(\beta)\pi_\mu(\mu).$$

The distributions $p(\mathbf{z}|\beta')$ and $p(\mathbf{x}|\beta)$ are identical, both involving the same intractable normalizing constant. Markov chain Monte Carlo simulations may proceed by the following in place of Step 3, the update for β, of Section 2.2. First, perform a block Gibbs update

for $(\beta', \mathbf{z})$, drawing β' from $p(\beta'|\beta)$, then drawing $\mathbf{z}$ from $q(\mathbf{z}|\beta')$ (by perfect simulation). Second, propose a swap of β with β', in which the intractable normalizing constants cancel from the following Metropolis–Hastings ratio:

$$\frac{q(\mathbf{z}|\beta)p(\beta|\beta')q(\mathbf{x}|\beta')\pi_\beta(\beta')}{q(\mathbf{z}|\beta')p(\beta'|\beta)q(\mathbf{x}|\beta)\pi_\beta(\beta)}.$$

5.1.3 Results From Exact Method

We have found that the Murray exchange algorithm appears to converge more quickly than the auxiliary variable method with simple auxiliary variable distributions. In addition, it has the advantage of not requiring an auxiliary distribution which approximates the intractable model. Computation for each update is comparable for the two methods, dominated by the perfect simulation step. We therefore present results based on the exchange algorithm. In general, our perfect sampling algorithm used to sample $\mathbf{z}$ from $p(\mathbf{z}|\beta')$ took an increasingly long time as β_1 increased. For example, for the data generated from the autologistic model with $\beta = (0, 0.3)$ and $\mu = (-0.5, 0.5)$, the auxiliary variable method took 5 days to carry out inference for 50 datasets, where 10,000 MCMC iterations were used for each dataset. The results are presented in Table 3. Comparing these results to those for the approximate methods in Table 1, we see that both the RDA and block pseudo-likelihood methods perform remarkably similarly, on average. However, we found that our perfect sampling algorithm slowed down considerably when the hidden autologistic parameters took the values $\beta = (0, 0.4)$. A single dataset took approximately 12 hr to compute. Therefore to carry out inference for all 50 datasets for each of the four different parameter configurations involving $\beta = 0.4$ would take approximately 100 days. However, we can report that on a small subset of the simulated datasets, again, our introduced approximate methods performed very similarly to the auxiliary variable method, but at a fraction of the computing time.

6. REAL DATA EXAMPLE

6.1 HIDDEN GENOMIC INTERACTIONS

Microarray technology allows simultaneous measurement of many gene expression levels. In a recent experiment (Bozdech et al. 2004), gene expressions were measured across the whole genome of *Plasmodium falciparum*, the organism that causes human malaria, for

Table 3. Average absolute bias of the posterior mean (and standard error of the posterior mean) for 50 samples of 50×50 lattices, using the exchange method, where the underlying hidden MRF is an autologistic model with $\beta = (0, 0.3)$ and $\mu = (-0.5, 0.5)$.

β_0	β_1	μ_{-1}	μ_1	True values
0.026 (0.005)	0.031 (0.002)	0.066 (0.013)	0.074 (0.003)	$\beta = (0, 0.3)$
0.032 (0.006)	0.032 (0.003)	0.095 (0.010)	0.055 (0.002)	$\beta = (0.05, 0.3)$

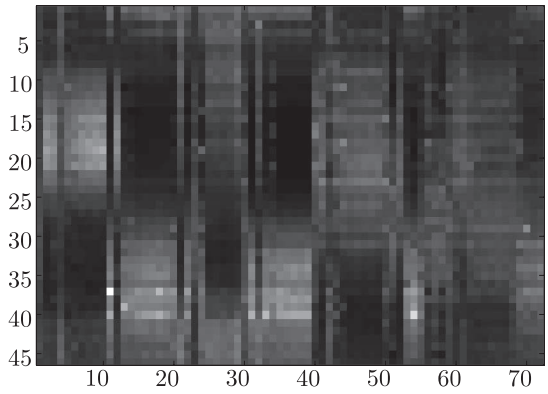

Figure 3. Log-differential expression levels for the mitochondrial genome across 46 1-hr time intervals. Columns are genes and rows are time points.

46 1-hr consecutive intervals. The experiment was conducted over the complete asexual intraerythrocytic development cycle to establish which genes might be potential drug targets for deregulating the organism to prevent malaria. The *Plasmodium falciparum* genome consists of 14 linear chromosomes, a circular genome, and a linear mitochondrial genome. In this example, we focus on the relatively short mitochondrial chromosome, which consists of 72 genes and about which relatively little is known.

We define the observations on a 46×72 spatial-temporal rectangular lattice where y_{tg} is the log-expression of the gth gene at time point t. Figure 3 displays the data $\mathbf{y}$.

From a biological point of view, it is interesting to model whether genes are down- or up-regulated and whether this pattern shows any spatial structure. The original publication (Bozdech et al. 2004) suggested that there was little evidence for spatial coregulation except on the circular genome, but they used, rather crudely, ordinary Pearson correlations on the original log-expressions. We investigate the temporal-spatial structure of the interaction between up- and down-regulated states within the mitochondrial chromosome of the *Plasmodium falciparum*, using the methods presented in this article. For this example we assume that the data hide a lattice of latent states $\mathbf{x}$ modeled as a nonhomogeneous autologistic distribution with two states $\{-1, 1\}$ corresponding to 'up-regulation' and 'down-regulation.' Thus the likelihood of $\mathbf{x}$ given model parameters β appears as

$$p(\mathbf{x}|\beta) \propto \exp(\beta_0 V_0(\mathbf{x}) + \beta_t V_t(\mathbf{x}) + \beta_g V_g(\mathbf{x})), \tag{6.1}$$

where $V_t(\mathbf{x})$ measures the interactions between neighboring lattice points corresponding to the same gene in the 'time' direction, whereas $V_g(\mathbf{x})$ similarly measures interactions at the same time point between neighboring genes. The parameters β_t and β_g allow for the possibility that the strength of the interaction might not be the same in both directions. The parameter β_0, as before, controls the relative abundance of each state. Of course other models could also be proposed to capture more information, for instance, extending this model to a three-state model including a state of 'no differential expression.' However, Bozdech et al. (2004) suggested that a vast majority of the genes are active, and so we ignore this possibility here. Further extensions might include time- or gene-specific interaction parameters.

Table 4. Posterior means (and standard deviations) of model parameters.

	β_0	β_t	β_g	μ_-	μ_+	σ
RDA	−0.009	1.429	0.159	0.812	2.060	0.509
	(0.003)	(0.025)	(0.015)	(0.016)	(0.040)	(0.010)
Block pseudo	−0.005	1.334	0.12	0.806	2.064	0.503
	(0.004)	(0.075)	(0.02)	(0.017)	(0.039)	(0.009)
Pseudo	0.048	1.370	1.252	0.933	1.963	0.627
	(0.189)	(0.262)	(0.382)	(0.068)	(0.282)	(0.048)

Returning to the current example—the distribution of $\mathbf{y}$ given $\mathbf{x}$ is modeled as independent Gaussian noise, with a fixed mean μ_- or μ_+ conditional on the corresponding state variable equal to -1 or $+1$, respectively, with common variance σ^2. The assumption of normality of log-expression levels has been shown to be reasonable for similar experimental conditions (Wit and McClure 2004). Inference was carried out by updating all parameter values from their full conditional distributions. Flat Gaussian zero-mean priors were chosen for each of the β parameters. A diffuse inverse Gamma prior was specified for σ. Prior values for μ were distributed uniformly from the set $\{(\mu_{-1}, \mu_1) | -5 \leq \mu_{-1} \leq 5, \mu_{-1} \leq \mu_1 \leq 5\}$. The boundary values, $(-5, 5)$, were chosen on the basis that log-expression levels for similar experiments were considerably inside this range, yielding an uninformative prior. Finally $\mathbf{x}$ was updated via Gibbs sampling for each lattice point x_i.

Posterior mean (and standard deviations) of model parameters are given in Table 4.

Comparing the posterior mean and standard deviations of parameters from each of the three approximation methods, the same patterns emerge as those of the simulation study. Namely, both the RDA and block pseudolikelihood methods give more precise estimates for the β parameters than the pseudolikelihood method. As far as inference for μ is concerned, again both the RDA and block pseudolikelihood methods give quite precise estimates, relative to the pseudolikelihood method. These results, in light of the evidence presented in the simulation study, suggest that the RDA and block pseudolikelihood methods yield useful estimators of model parameters.

The large value for β_t shows that it is mainly persistent time-effects that are responsible for the structured pattern in the data. However, there is also a significantly positive gene effect, indicated by the positive values of β_g, which suggests that the change in expression of a gene tends to coincide with a change in the same direction of the two neighboring genes. Most likely, this positive spatial effect is due to the operon structure. A transcription factor may bind upstream from several genes and may be responsible for expressing all genes in that region.

In Figure 4 an image of the marginal posterior probabilities of a particular lattice point belonging to state 1 is given, where RDA is used in the likelihood approximation. From this lattice, a reconstruction of $\mathbf{x}$ is derived by thresholding each lattice point at 0.5 probability. By looking across the genome direction, this image can be used to determine how many transcription factors control the expression of this chromosome. It seems that the expressions of these 76 genes are controlled by at least two transcription factors. In the time direction, the reconstruction shows clearly that the expression spans exactly one cell-cycle.

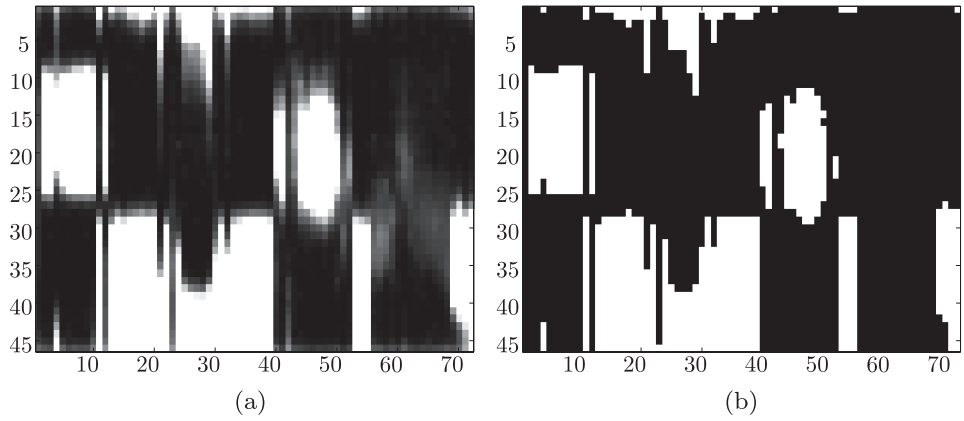

Figure 4. (a) An image displaying posterior probabilities that each lattice point takes the value +1. Dark intensities indicate low probability, whereas light indicates high probability. (b) A thresholded version of image (a) (at threshold probability 0.5).

7. DISCUSSION

In this article we have evaluated several approximations to the likelihood of a realization from a binary Markov random field in the context of inference for hidden Markov random fields using Markov chain Monte Carlo procedures. In particular, our reduced dependence approximation appears to be superior to the other methods investigated, in terms of posterior bias and computational efficiency. It is also to be preferred because it implies multiplication of the prior probability by a constant factor in each iteration of the Markov chain, which ensures positive recurrence.

The extension of our likelihood approximations to the Potts model, where the number of latent states is more than two, is straightforward. However, their dependence on the forward recursion method would place an upper limit on the number of states which can be practically accommodated in a reasonable time, and a consequent reduction in the number of rows m_1 in the reduced dependence approximation, and the number of rows in blocks for the block pseudolikelihood method.

Finally, the methods we propose have been shown to be useful in the context of a gene expression dataset, successfully quantifying time and gene neighboring effects. We expect similar datasets displaying two-dimensional lattice structure to benefit in their modeling from the use of our methods.

SUPPLEMENTAL MATERIALS

C code: The supplemental files for this article include C programs which can be used to replicate the simulation study in this article. Please see the contained file `README.txt` for more details. (Friel-etal-simulation.zip, zip archive)

ACKNOWLEDGMENTS

We thank the editors and referees for their helpful comments which improved the article. The research of the first three authors was supported by the Australian Research Council.

[Received November 2006. Revised March 2009.]

REFERENCES

Abend, K., Harley, T., and Kanal, L. (1965), "Classification of Binary Random Patterns," *IEEE Transactions on Information Theory*, IT-11, 538–544.

Augustin, N., Mugglestone, M., and Buckland, S. (1996), "An Autologistic Model for Spatial Distribution of Wildlife," *Journal of Applied Ecology*, 33, 339–347.

Besag, J. E. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 192–236.

——— (1975), "Statistical Analysis of Non-Lattice Data," *The Statistician*, 24, 179–195.

Besag, J., York, J., and Mollié, A. (1991), "Bayesian Image Restoration, With Two Applications in Spatial Statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu J., and DeRisi, J. L. (2004), "The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium Falciparum," *PLoS Biology*, 1 (1), 85–100.

Cox, D. R., and Reid, N. (2004), "A Note on Pseudolikelihood Constructed From Marginal Densities," *Biometrika*, 91 (3), 729–737.

Cressie, N., and Davidson, J. (1998), "Image Analysis With Partially Ordered Markov Models," *Computational Statistics and Data Analysis*, 29 (1), 1–26.

Dormann, C. F. (2007), "Assessing the Validity of Autologistic Regression," *Ecological Modelling*, 207, 234–242.

Dryden, I. L., Scarr, M. R., and Taylor, C. C. (2003), "Bayesian Texture Segmentation of Weed and Crop Images Using Reversible Jump Markov Chain Monte Carlo Methods," *Applied Statistics*, 52 (1), 31–50.

Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Contants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 657–699.

Goutsias, J. (1991), "Unilateral Approximation of Gibbs Random Field Images," *Computational Vision Graphics and Image Processing: Graphical Models and Image Processing*, 53, 240–257.

Green, P. J., and Richardson, S. (2002), "Hidden Markov Models and Disease Mapping," *Journal of the American Statistical Association*, 97, 1055–1070.

Gu, M. G., and Zhu, H.-T. (2001), "Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation," *Journal of the Royal Statistical Society*, Ser. B, 63, 339–355.

Heagerty, P. J., and Lele, S. R. (1998), "A Composite Likelihood Approach to Binary Spatial Data," *Journal of the American Statistical Association*, 93, 1099–1111.

Heikkinen, J., and Högmander, H. (1994), "Fully Bayesian Approach to Image Restoration With an Application in Biogeography," *Applied Statistics*, 43, 569–582.

Huang, F., and Ogata, Y. (2002), "Generalized Pseudo-Likelihood Estimates for Markov Random Fields on Lattice," *Annals of the Institute of Statistical Mathematics*, 54 (1), 1–18.

Jordan, M. (2004), "Graphical Models," *Statistical Science*, 19, 140–155.

Liang, F. (2007), "Continuous Contour Monte Carlo for Marginal Density Estimation With an Application to a Spatial Statistical Model," *Journal of Computational and Graphical Statistics*, 16 (3), 608–632.

McGrory, C. A., Titterington, D. M., Reeves, R., and Pettitt, A. N. (2008), "Variational Bayes for Hidden Markov Random Field Analysis," *Statistics and Computing*, DOI 10.1007/s11222-008-9095-6.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006), "An Efficient Markov Chain Monte Carlo Method for Distributions With Intractable Normalising Constants," *Biometrika*, 93, 451–458.

Murray, I. (2007), "Advances in Markov Chain Monte Carlo Methods," Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, U.K.

Murray, I., and Ghahramani, Z. (2004), "Bayesian Learning in Undirected Graphical Models: Approximate MCMC Algorithms," in *Uncertainty in Artificial Intelligence (UIA-2004)*, pp. 392–399.

Pettitt, A. N., Friel, N., and Reeves, R. (2003), "Efficient Calculation of the Normalising Constant of the Autologistic and Related Models on the Cylinder and Lattice," *Journal of the Royal Statistical Society*, Ser. B, 65 (1), 235–247.

Preisler, H. K. (1993), "Modelling Spatial Patterns of Trees Attacked by Bark-Beetles," *Applied Statistics*, 42, 501–514.

Propp, J. G., and Wilson, D. B. (1996), "Exactly Sampling With Coupled Markov Chains and Applications to Statistical Mechanics," *Random Structures and Algorithms*, 9, 223–252.

Reeves, R., and Pettitt, A. N. (2004), "Efficient Recursions for General Factorisable Models," *Biometrika*, 91, 751–757.

Rydén, T., and Titterington, D. M. (1998), "Computational Bayesian Analysis of Hidden Markov Models," *Journal of Computational and Graphical Statistics*, 7, 194–211.

Scott, S. L. (2002), "Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century," *Journal of the American Statistical Association*, 97 (457), 337–351.

Sebastiani, G., and Sørbye, S. H. (2002), "A Bayesian Method for Multispectral Image Data Classification," *Nonparametric Statistics*, 14, 169–180.

Sherman, M., Apanasovich, T. V., and Carroll, R. J. (2006), "On Estimation in Binary Autologistic Spatial Models," *Journal of Statistical Computation and Simulation*, 76, 167–179.

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Ser. B, 66 (2), 275–296.

Wit, E., and McClure, J. (2004), "Statistics for Microarrays: Design, Analysis and Inference," Chichester: Wiley.

Wu, H., and Huffer, F. W. (1997), "Modelling the Distribution of Plant Species Using the Autologistic Regression Model," *Ecological Statistics*, 4, 49–64.

Zhu, H. T., Gu, M. G., and Peterson, B. (2007), "Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation," *Statistics and Computing*, 17, 163–177.

Zucchini, W., and Guttorp, P. (1991), "A Hidden Markov Model for Space Time Precipitation," *Water Resources Reseach*, 27, 1917–1923.